2016

An unconstrained statistical matching algorithm for combing individual and household level geo-specific census and survey data

Mohammad-Reza Namazi-Rad
University of Wollongong

Robert Tanton
University of Canberra

David Steel
University of Wollongong

Payam Mokhtarian
Fairfax Media Limited

Sumonkanti Das
University of Wollongong

An unconstrained statistical matching algorithm for combing individual and household level geo-specific census and survey data

Abstract

The Population Census is an important source of statistical information in most countries that is capable of producing reliable estimates of population characteristics for small geographic areas. One limitation of a census is that there are many population characteristics that cannot be collected due to respondent burden or cost. This means that statistical agencies have to conduct population based surveys to provide social, economic and demographic characteristics for a target population which are not captured by a large-scale census. These surveys are usually capable of producing direct estimates at the national level and high level regions but often cannot produce reliable estimates for smaller areas. Due to the increasing demand for comprehensive statistical information not only at the national level but also for sub-national domains, there is a wide discussion in the literature about the use of statistical techniques that combine survey with census data to provide more detailed, finer-level estimates.

Where censuses and sample surveys are based on the same reporting units, statistical matching techniques can be employed to link the records from survey and census data where exact matching of reporting units is impossible due to confidentiality restrictions. These techniques can then provide the detailed social, economic and demographic information required for small areas.

An approach is developed in this paper in which a close-to-reality synthetic population of individuals and households is generated from available census tables using an iterative proportional updating (IPU) method. Statistical matching using a nearest neighbour method is then used to impute survey data to the individuals and households in the synthetic population. To evaluate this approach, 2011 Bangladesh census data is used to generate a district-specific synthetic population of individuals and households. Matching is then performed by imputing the nearest possible records among the 2011 Bangladesh Demographic and Health Survey to estimate the wealth index for each household within the synthetic population. The results show that using the method presented in this paper helps with achieving more representative estimates (comparing with direct survey estimates,) particularly for areas with small sample sizes where not all population units with different socio-demographic characteristics are included.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

01-16

An Unconstrained Statistical Matching Algorithm for Combing Individual and Household Level Geo-Specific Census and Survey Data

Mohammad-Reza Namazi-Rad, Robert Tanton, David Steel, Payam Mokhtarian, Sumonkanti Das

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

An Unconstrained Statistical Matching Algorithm for Combining Individual and Household Level Geo-Specific Census and Survey Data

Mohammad-Reza Namazi-Rad[a,*], Robert Tanton[b], David Steel[a], Payam Mokhtarian[c], Sumonkanti Das[a]

[a] *National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, AUSTRALIA*
[b] *National Centre for Social and Economic Modelling, University of Canberra, ACT 2601, AUSTRALIA*
[c] *Fairfax Media Limited, NSW 2009, AUSTRALIA*

Abstract

The Population Census is an important source of statistical information in most countries that is capable of producing reliable estimates of population characteristics for small geographic areas. One limitation of a census is that there are many population characteristics that cannot be collected due to respondent burden or cost. This means that statistical agencies have to conduct population based surveys to provide social, economic and demographic characteristics for a target population which are not captured by a large-scale census. These surveys are usually capable of producing direct estimates at the national level and high level regions but often cannot produce reliable estimates for smaller areas. Due to the increasing demand for comprehensive statistical information not only at the national level but also for sub-national domains, there is a wide discussion in the literature about the use of statistical techniques that combine survey with census data to provide more detailed, finer-level estimates.

Where censuses and sample surveys are based on the same reporting units, statistical matching techniques can be employed to link the records from survey and census data where exact matching of reporting units is impossible due to confidentiality restrictions. These techniques can then provide the detailed social, economic and demographic information required for small areas.

An approach is developed in this paper in which a *close-to-reality* synthetic population of individuals and households is generated from available census tables using an iterative proportional updating (IPU) method. Statistical matching using a nearest neighbour method is then used to impute survey data to the individuals and households in the synthetic population. To evaluate this approach, 2011 Bangladesh census data is used to generate a district-specific synthetic population of individuals and households. Matching is then performed by imputing the nearest possible records among the 2011 Bangladesh Demographic and Health Survey to estimate the wealth index for each household within the synthetic population. The results show that using the method presented in this paper helps with achieving more representative estimates (comparing with direct survey estimates,) particularly for areas with small sample sizes where not all population units with different socio-demographic characteristics are included.

Keywords: Imputation; Spatial Microsimulation; K-Nearest Neighbours; Pseudo Census; Small Area Estimation; Synthetic Population.

☆

*Corresponding author:Mohammad-Reza Namazi-Rad
Email: mrad@uow.edu.au

1. Introduction

The need for reliable and accurate information concerning poverty, inequality, and living conditions of people and households for geographic areas has increased substantially in recent years. Such information is a basic instrument for targeting policies and programs aimed at the reduction of poverty. Household surveys collect information on incomes, expenditures, and demographics to generate estimates of wealth and poverty at a national level and possibly for large geographic areas a country. However, data confidentiality conditions generally restricts access to unit level survey data with small area identifiers. Even if access to such data is possible, the small sample sizes result in unreliable direct estimates for small areas. This is mostly the case in developing countries. Therefore, indirect estimation approaches are employed for area-level poverty mapping in different parts of the world; e.g. South Africa (Alderman et al. (2002)), Ecuador (Elbers et al. (2003)), Mexico (Tarozzi & Deaton (2009)), India (Coondoo et al. (2011)), and Spain (Molina & Rao (2010)). A review of such methods is presented by Chambers & Pratesi (2014). Here, a micro-simulation technique is presented for measuring the area-specific wealth indices in different parts of Bangladesh.

In terms of estimation, *small areas* are the geographic or demographic subsets of the population whose domain-specific sample size is not large enough to produce reliable direct estimates. *Large areas* are those with enough domain-specific sample information to warrant the use of direct estimators solely based on data obtained from that area. During the last few decades, different small area estimation (SAE) techniques have been developed to overcome the challenging problem of finding reliable estimates for small areas (e.g. Rao (2003), Chambers & Tzavidis (2006), Chambers et al. (2014), Namazi-Rad & Steel (2015), Chandra et al. (2015)). In particular, spatial microsimulation techniques are increasingly used to derive small area estimates of many indicators using survey data (Tanton et al. (2009), Tanton et al. (2011), Tanton & Clarke (2014), Burden & Steel (2015)).

Spatial-microsimulation models are being used for synthesising spatial micro data based on real input data. Such methods are increasingly used to model the behaviour of individual entities or agents in different applications. Microsimulation has become commonplace in generating a spatially disaggregated population micro-dataset (Ballas et al. (2006) , Morrissey et al. (2014), Farrell et al. (2013), Morrissey et al. (2008)), modelling population aging and household transitions (Namazi-Rad et al. (2014a), Namazi-Rad et al. (2014b)), and dynamics of regional and local labour markets (Morrissey et al. (2008), Farrell et al. (2013)). The approach is being increasingly used in modelling the economy (Kokic et al. (2000), Morrissey et al. (2014)), urban energy markets (Mozumder & Marathe (2005)), education (Wu et al. (2008)), agri-environment (Hynes et al. (2008), Hynes et al. (2009)), policy making (e.g. Lovelace & Ballas (2013)), public health (Tomintz et al. (2008), Edwards et al. (2011)), population movements and traffic analysis (Lovelace et al. (2014), Treiber & Kesting (2013)), and disease control (Eubank et al. (2004), Barrett et al. (2005), Ferguson et al. (2006)). Details about standard microsimulation models are discussed by Wu et al. (2008), Anderson & Hicks (2011), Birkin & Clarke (2011), and O'Donoghue (2015).

In this paper, a hybrid spatial microsimulation technique is presented for generating an area-specific synthetic population (SP) of individuals and households which will be considered as a pseudo-census for the purpose of the current study. By using this novel approach, the theory behind both sample-based synthesis approaches (as discussed by Wilson & Pownall (1976), Arentze et al. (2007), Guo & Bhat (2007), and Namazi-Rad et al. (2014a)) and sample-free population synthesis approaches (as discussed by Voas & Williamson (2001) are considered to achieve more accurate area-specific population estimates. To do so, an artificial population is to be simulated from anonymous census data at the individual and household levels which realistically matches the observed population in a geographical area for a given set of of table margins. Using this approach, the identification of population units and/or their sensitive information in the generated area-specific synthetic data will be difficult (Beckman et al.

(1996), Rubin (1987)).

Once the reliable SP is generated, survey-based estimates can be projected over the entire population using the statistical matching techniques based on the same reporting units. For measuring population-specific indicators based on available census and sample data, and more specifically for measuring the poverty and wealth indicators, which is of the main concern in the current study, having close-to-reality population data helps to create a more accurate imputation of survey data based on population characteristics correctly classified within the SP. For empirical evaluation the Bangladesh census data is used to generate a district-specific SP of individuals and households. Statistical matching is then performed by imputing the nearest possible records among the 2011 Bangladesh Demographic and Health Survey (2011 BDHS) to calculate synthesised wealth indicators for the entire population at the level of households. The wealth index is calculated for each survey individual in 2011 BDHS based on a method developed at the Bangladesh Bureau of Statistics (BBS). This method is briefly discussed in this paper.

2. Population Synthesis

A synthetic population aims at faithfully reproducing actual social entities, such as individuals and households, and their characteristics as described in a population census. Depending on the quality and completeness of the input datasets, as well as the number of variables of interest and hierarchical levels (usually, individual and household), a reliable SP should be able to reflect the actual physical social entities, with their characteristics and specific behavioural patterns (Namazi-Rad et al. (2014b)).

The synthetic reconstruction (SR) approach has been traditionally used by researchers (e.g. Namazi-Rad et al. (2014a); Farooq et al. (2013)) for generating SP using both disaggregated- and aggregated-level data. This method first uses available disaggregated-level data while assuming that it is a representative sample of the target population. This is generally referred to as the *seed*

data. Then, population units with the required socio-demographics are randomly drawn from the representative disaggregated-level data and populated within the target area using a weighting technique so that the marginal distribution follows the aggregated-level information coming from one source covering the complete population (e.g. census data).

In order to generate a reliable SP, multi-dimensional tables of population units' socio-demographic variables are needed. When dealing with area-specific tables at the lower dimension, the iterative proportional fitting procedure (IPFP) is proposed by Deming & Stephan (1940), as an algorithm that adjusts a table of data in a way that table cells add up to given totals in all required dimensions. This application of IPFP to contingency tables with known margins is called *raking* and are discussed by Deming & Stephan (1940); Stephan (1942); Fienberg (1970); Deville et al. (1991); Lu & Gelman (2003); Namazi-Rad et al. (2014a). The iterative proportional updating (IPU) is also proposed by Ye et al. (2009) and Pritchard & Miller (2012), as a hierarchical IPFP, to control household and person level attributes, simultaneously. In other terms, the IPU algorithm is employed to estimate sample household weights that satisfy both household and person type constraints. Once the multi-dimensional tables are generated, the seed data is to be used together with these tables in the SR approach to reconstruct the population individuals and households with a computer-based simulation.

A practical issue with the sample-based approach is that the representativeness of the resulting SP is highly dependent on how well the seed data is representative of the entire population. When dealing with multi-dimensional census tables, and in particular when using individual-specific and household-specific tables simultaneously, it is often hard to find representative units in the seed data associated with all cells in the census tables. The alternative to sample-based approaches are sample-free approaches, discussed in the literature for population synthesis where a representative sample is not available (Gargiulo et al. (2010), Lenormand & Deffuant (2013)). Sample-free approaches generally do not rely on a sample record file to construct a SP. Instead they rely on

heuristics to ensure that the geographical heterogeneity of the resulting SP is best preserved. One major issue with the sample-free approaches is that the source of variability in the simulated SP is not clearly identified.

A hybrid approach is presented here which starts with a sample-based algorithm for population synthesis using census tables and a sample record file. Where dealing with the population cross-classified counts for which representative units are missing in the sample data file, a sample-free algorithm is employed. Using the approach presented in this section an area-specific SP is then constructed for Bangladesh using its 2011 census data.

2.1. Notations and Methodology

This paper addresses the structural hierarchies in developing a SP in which household structures and socio-demographics of individuals living within households are considered. A certain number of characteristics (denoted by P) are considered to define any specific individual within the SP. Then, SP individuals form single-member and multi-member households, each with certain number of characteristics (denoted by Q). Given a finite set of characteristics P for individuals and a finite set of characteristics Q for households, the state model here is defined as an onto (surjective) function from P to Q ($f : \{1, 2, \ldots, P\} \rightarrow \{1, 2, \ldots, Q\}$) satisfying certain additional constraints, where $|P| < |Q|$. The study area is assumed to be divided to G overall areas and each household (with all its individuals) is located at one specific area; i.e. $g \in \{1, 2, \ldots, G\}$.

We consider a population $\mathbb{U} = (\mathbb{U}^{(H)}, \mathbb{U}^{(I)})$ of size N individuals where each population individual by itself or with several other individuals belong to uni- or multi-member household. The total number of population households is denoted by M. The superscript '(H)' refers to the households and '(I)' refers to the individuals. So, $\mathbb{U}^{(H)}$ refers to the population of households and $\mathbb{U}^{(I)}$ refers to the population of individuals. Assuming the target of inference to be at the area level, the whole population is divided into G areas (i.e. $\mathbb{U}_g = (\mathbb{U}_g^{(H)}, \mathbb{U}_g^{(I)})$), with M_g households and N_g individuals in the gth area, where $M = \sum_{g=1}^{G} M_g$

& $N = \sum_{g=1}^{G} N_g$. The jth household in the gth area is denoted by $Y_{jg} \in \mathbb{U}_g^{(H)}; j \in \{1, 2, \ldots, M_g\}$ and the ith individual in this household is denoted by $X_{ijg} \in \mathbb{U}_{jg}^{(I)}$.

In the sample-based population synthesis, a representative sample file $\mathbb{S}_g = (\mathbb{S}_g^{(H)}, \mathbb{S}_g^{(I)})$ is assumed to be available for the gth area with m_g households denoted by $y_{jg} \in \mathbb{S}_g^{(H)}; j \in \{1, 2, \ldots, m_g\}$ and n_g individuals, out of which those belonging to the jth household are denoted by $x_{ijg} \in \mathbb{S}_{jg}^{(I)}$. Here, $\mathcal{T}(X_{ijg})$ and $\mathcal{T}(Y_{jg})$ denote the vectors of attributes of population individuals and households, respectively. Additionally, $\mathcal{T}(x_{ijg})$, and $\mathcal{T}(y_{ig})$ denote the attributes of sample individuals and households, respectively.

One common known approach for estimation of population counts in multi-dimensional contingency tables when a random sample is available together with marginal population tables of lower dimensions is IPFP. For sample-based population synthesis, individual- and household-specific multi-dimensional cross-tabulation of population counts are estimated using the IPU algorithm as a hierarchical IPFP (presented by Lenormand & Deffuant (2013)), based on the representative sample data subject to known marginal population counts. IPU is used as an algorithm for estimating the household weights in a way that household and individual distributions are matched. Then, the SR approach is used as a deterministic algorithm for reconstructing the population.

Using the SR approach, individuals and households with the required socio-demographics are populated within each area using a weighting technique so that the marginal distribution follows the aggregated-level information coming from one source covering the complete population. One way to do this is to use the deterministic re-weighting algorithm (e.g. Ballas et al. (2005), Smith et al. (2009)) to allocate a weight to each unit record within the seed data (i.e. disaggregated-level data) and consider the weights as a distribution of probabilities derived from the available seed data. Each attribute for the population units is treated separately and sampling from marginal distributions is conducted to select the number of units equal to the number of area-specific population totals.

In practice, a perfect alignment between all area-specific synthetic totals and

the sample totals under the assumption that all of the areas are relatively homogeneous is unrealistic. To overcome this challenge, it is helpful to conduct a Monte Carlo sampling from the disaggregated data based on the underlying conditional probabilities calculated (as discussed by Harland et al. (2012)), rather than being deterministically re-weighted from the disaggregated-data. This is a stochastic approach as the conditional probabilities are readjusted using iterative Monte Carlo sampling until a close match with the constraining tables or marginal distributions is achieved. As alternative proposed here is to employ a hybrid approach using step-wise sample-based and a sample-free approach.

Let $\boldsymbol{\pi}_{g(I)}$ denote the vector of all cross-classified probabilities for the P individual-specific characteristics and $\boldsymbol{\pi}_{g(H)}$ denote the vector of all cross-classified probabilities for the Q household-specific characteristics. In the sample-based population synthesis, individuals and households from the seed data are allocated to the SP based on the population probabilities in $\boldsymbol{\pi} = \left(\boldsymbol{\pi}_{g(I)}, \boldsymbol{\pi}_{g(H)}\right)$ with the constraint of matching with the population margins. The pseudo code for this stage of population reconstruction for gth area is shown in Algorithm 1.

Algorithm 1 Population Synthesis Algorithm Stage I: Sample-Based Approach

Inputs:

(i) Individual-specific cross-classified population probabilities ($\pi_{g(I)}$)

(ii) Household-specific cross-classified population probabilities ($\pi_{g(H)}$)

(iii) Seed Data [t($\mathbf{x}_g$) , t($\mathbf{y}_g$)]

Output: Sample-based SP $\hat{\mathbb{U}}_g^{SB-SP}$

Algorithm:

1: **while** Changes in the resulted SP is not negligible over a sequence of iterations **do**

2: Pick at random household y_{jg} form the seed with all belonging individuals x_{ijg}

3: Find the associated population weights from vector of population probabilities; i.e. $\boldsymbol{\pi} = \left(\boldsymbol{\pi}_{g(I)}, \boldsymbol{\pi}_{g(H)}\right)$

4: Add the selected household and individuals to the SP $(y_{ig}, x_{ijg}) \xrightarrow{\boldsymbol{\pi}} (Y_{ig}, X_{ijg})$: subject to known marginal population counts

5: **end while**

The IPU approach used relocates weights among the census sample households of a type to account for differences in household composition and the individuals' characteristics constructing each household. To do so, the household-

specific weights have to be adjusted in a way to match the individual-specific constraints as closely as possible. In order to reconstruct the population of a specific area (say gth area), as presented in Algorithm 1, the resulting population weights/probabilities (i.e. $\boldsymbol{\pi} = \left(\boldsymbol{\pi}_{g(I)}, \boldsymbol{\pi}_{g(H)}\right)$) are used to pick the population units from the seed data and populate them based on the pre-identified weights.

Once the first stage in generating the SP is finished, it is time to compare the marginal counts in the simulated SP with population margins presented in the form of census tables. Where the marginal counts are not the same, it means that the seed data used for generating the synthetic population in the associated categories were not completely informative. To deal with this issue and to simulate the missing units in the SP once detected, we use a heuristic algorithm. Following Huynh et al. (2013), we need to construct a pool of individuals and a pool of households based on the multi-dimensional population cross-classifications obtained by employing IPFP. For the gth area, individual X_{ig}^P is characterized by a vector of attributes (i.e. $\mathcal{T}(X_{ig}^P)$) based on the individual-specific population cross-classifications for the gth area. The same method is to be employed for constructing household Y_{jg}^P characterized by a vector of attributes (i.e. $\mathcal{T}(Y_{jg}^P)$) based on the household-specific population cross-classifications for the gth area.

In the second stage of population synthesis, the missmatches between the population cross-classified counts in the SP simulated for the gth area and the census multi-dimensional contingency tables will inform the drawing process of individuals and the way the individuals are allocated to the households to solve the issues. To construct the households while the representative units are missing, records of individuals are drawn from the pool of individuals and allocated into households so that the resulting households in the SP satisfy the desired joint distributions at household level. This is achieved while preserving the distribution computed at the individual level. The joint distributions at individual level also informs this drawing process in terms of the probability an individual type being drawn given the household type being considered and attributes of the existing (previously allocated) residents.

Following Lenormand & Deffuant (2013), the list of individuals are located in each household one by one. The individuals are selected one at a time by order of importance in the household conditioned an the attributes of the previous individuals picked for this household. Once a household is built, it is added to the SP. The pseudo code for this stage of population synthesis of the gth area is shown in Algorithm 2. The first step in this Algorithm 2 is to identify the cross-classified probabilities calculated for the SP generated using Algorithm 1 which does not match with those in the original census tables and not those obtained using IPU. Once the mismatches are recognised, the associated population units will be swapped with new units in way to correct the mismatches, as much as possible.

Algorithm 2 Population Synthesis Algorithm Stage II: Sample-Free Approach

Inputs:

(i) Individual-specific cross-classified population probabilities ($\pi_{g(I)}$)

(ii) Household-specific cross-classified population probabilities ($\pi_{g(H)}$)

(iii) Pool of households

(iv) Pool of individuals

(v) Sample-based SP $\hat{\mathbb{U}}_g^{SB-SP}$

Output: Synthetic population of individuals and households $\hat{\mathbb{U}}_g$

Algorithm:

1: **for** Any missmatch between the household-specific cross-classified probabilities form the census tables (i.e. $\pi_{g(H)}$) and those calculated from the sample-based SP (i.e. $\hat{\pi}_{g(H)}^{SB-SP}$) **do**

2: Pick at random relevant household Y_{jg}^P form the pool of households

3: Pick at random individuals X_{ig}^Ps form the pool of individuals and locate them in the Y_{jg}^P considering $\mathcal{T}(Y_{jg}^P)$ and the constrains defined for a set of individuals to build a household

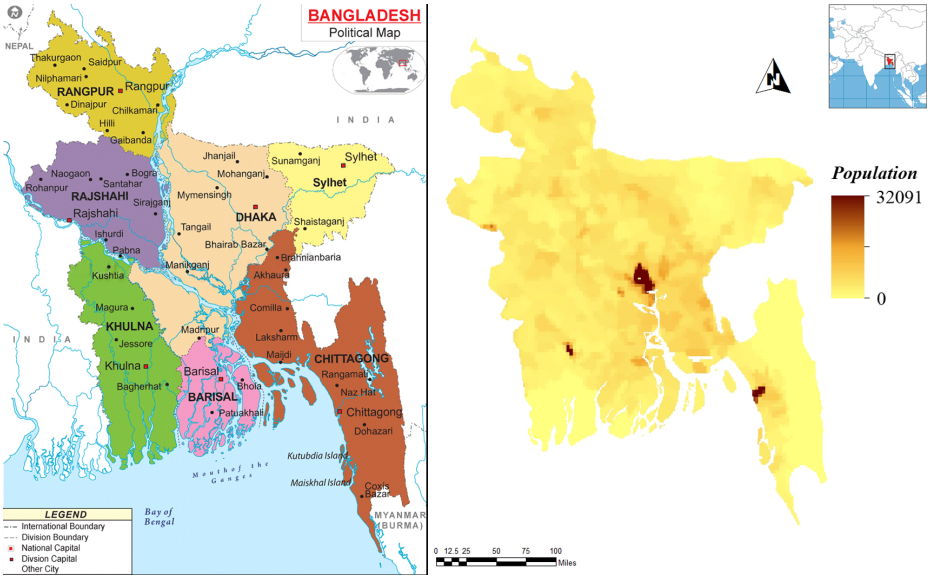
4: **end for**

2.2. Synthetic Population of Bangladesh

Bangladesh is the eighth most populous country with a population of size about 150 million and is the twelfth most densely populated country in the world (with 1015 individuals per square kilometre), according to the Bangladesh 2011 Census of Population and Housing. Bangladesh is located in the north-eastern part of South Asia and is mostly surrounded by the Indian borders. The country consists of 7 divisions, 64 districts (zila), and 545 sub-districts (upazila) with a majority of rural areas. A map of Bangladesh together with the population

maps is presented in Figure 1.

Figure 1: Bangladesh political map (source: mapsofworld.com) and population maps (number of persons per square kilometer of land area)



In the current study, we aim to simulate the district-specific SP of Bangladesh. For each district, the 2011 census data is available in the the form of one- and two-dimensional tables. The 5% Census Sample File (CSF) of households and individuals is also available. The SP in this case study consists of individuals defined by their age, gender, marital status and education while each individual (by him/herself or together with some other individuals) forms a household with a certain household type, household size and dwelling structure. The categorical variables for individuals and households to be simulated in the SP are presented in Table 1. In the sample-based approach, individual- and household-specific characteristics are obtained from the 5% CSF based on the 2011 Bangladesh census while being consistent with the area-specific census tables.

Table 1: Population attributes considered in population synthesis

Gender	**Age Category**	**Education**	**Marital Status**
(1) Male (2) Female	(1) 0-4 (2) 5-9 (3) 10-14 (4) 15-19 (5) 20-24 (6) 25-29 (7) 30-49 (8) 50-59 (9) 60-64 (10) 65+	(1) Literate (2) Illiterate	(1) Never married (2) Married (3) Widowed (4) Divorced or Separated

Household Type	**Dwelling Structure**	**Household Size**
(1) Single person household (2) Couple only (3) Couple with kids younger than 15 yrs (4) Couple with dependents over 15 yrs (5) Single parent with kids younger than 15 yrs (6) Single parent with dependents over 15 yrs (7) Other	(1) Pucka (Made in brick) (2) Semi-pucka (Made in brick, Tin, and Wood) (3) Kutcha (Built in wood/Bamboo) (4) Jhupri (Made in bamboo, leaves, Polythin, Others, etc)	(1) 1 (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8+

The population synthesis approach previously discussed in this section using Algorithm 1 and Algorithm 2, respectively, is used in simulating district-specific SP for which the 2011 Bangladesh census tabulated data corresponding to area counts are used as the know population margins. The 5% CSF is used as the seed data for population synthesis. This data covers 64 districts (Zila) and 544 sub-districts (Upzila) that represent the whole country and is provided by the BBS. The rural-urban classification by BBS for the population of each district is followed in this study when using the individual- and household-level information.

To evaluate the district-specific SP, we calculate the absolute relative bias to compare the cross-classified counts in the SP with those in the multi-dimensional tables archived when using IPU. Where there are I cross classifications, the ith cross-tabulated population probability calculated solely out of the census tables using the IPU approach are denoted by $\theta_i^{(IPU)}; i \in \{1, 2, \ldots, I\}$. The estimated

value for the same cross-tabulated population probability calculated based on the SP simulation is denoted by $\theta_i^{(SP)}$. Then, the absolute relative bias is calculated as follows:

$$\text{ARB}(\theta_i) = \left| \frac{\theta_i^{(SP)} - \theta_i^{(IPU)}}{\theta_i^{(SP)}} \right|$$

For the attributes presented in Table 1, Figure 2 and Figure 3 present the results for individuals- and household-level classification, respectively. The graphs illustrate the absolute relative bias for each cell so as to be ascending for the area-specific total population of districts. This means the 1st district in the graph is the one with the largest population size and the 64th district is the one with the smallest size of population. Based on the 2011 Bangladesh census, the largest district in terms of population size is Dhaka with population of 11,996,728 individuals and the smallest district in terms of population is Bandarbon with the population of 387,129 individuals. The populations of all 64 districts are discussed in more details in Section 3. These results presented in Figure 2 and 3 show that the construction of SP of individuals and households in Bangladesh is with relatively small error (i.er. the average absolute relative bias is less than 10%).

Figure 2: Absolute relative bias for individual-specific classes of attributes when comparing the cross-classified counts in the SP and census tables for each distinct (Stratum: 18 individual-specific population attributes as presented in Table 1; Districts: 64 Bangladesh Districts)

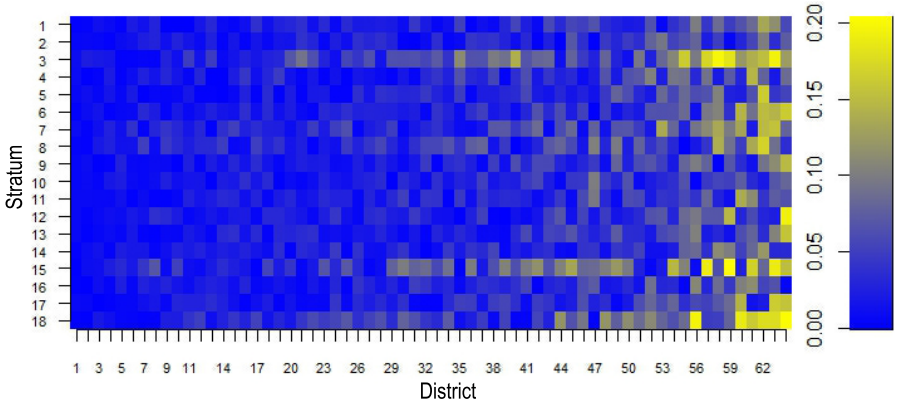
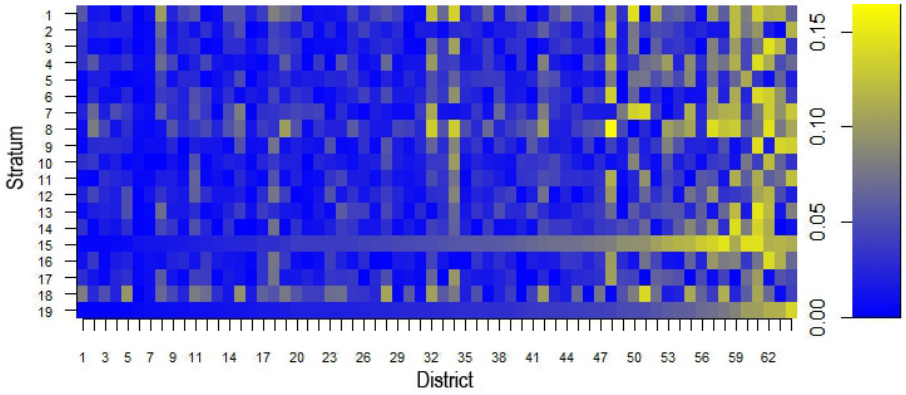


Figure 3: Absolute relative bias for household-specific classes of attributes when comparing the cross-classified counts in the SP and census tables for each distinct (Stratum: 19 household-specific population attributes as presented in Table 1; Districts: 64 Bangladesh Districts)



As shown in Figure 2 and Figure 3, cross-classified counts for different classes of attributes are closer to the actual population counts for districts with larger population sizes. For smaller districts (in terms of population size), finding exact matches has been more challenging. Additionally, errors for certain classes of attributes tend to be larger. For example, allocating individuals aged between 0 to 5 to the right households is more challenging, particularly in areas with small population size where an exact match in the seed data is not available or is hard to find. Another example is the households categorized as "Single Person" and "Other", for which a exact match is hardly obtained using the seed data.

3. Statistical Matching

Statistical matching is a statistical approach employed to provide information on the joint distribution of variables and indicators collected through two or more sources on the same population. It offers the possibility of increasing the value of the current data, without increasing costs and response burden. Statistical matching techniques aim to integrate two or more data sources referring to the same target population in cases when exact matching of individual records (record linkage) is not possible due to confidentiality restrictions on the

data available (Rosenbaum (2002), Rubin (2006)). The two data sources have to share one or more variable. The main objective is to fill in (impute) the dataset chosen as the *recipient* with the values of the variables which are available only in the other dataset, the *donor* one. Imputation approaches are also used as a typical post-survey strategy to compensate for missing data (Groves & Couper (1998), Lago & Clark (2015)).

For each unit in the recipient dataset, the key aim is to search for similar entities in the donor dataset and impute the value of the variable(s) only available in the donor dataset. Statistical matching uses variables common to both datasets to identify similar records that can be linked in order to generate a new synthetic dataset that allows more flexible analysis than would be possible with the two separate datasets (Rassler (2002)). For this purpose, a distance function needs to be defined to calculate how similar the recipient and donor data are and/or provide the clustering as a similarity-based segmentation.

Multi-variate and propensity score statistical matching techniques are discussed in the literature for generating suitable control groups that are similar to treated groups when a randomised experiment is not available (Rosenbaum & Rubin (1983), Austin (2011), An (2010)). Following the terminology presented by Rosenbaum & Rubin (1983), the propensity score is referred to as a balancing score defined as the treatment assignment. This method is used in other applications for pairing purposes Longford (2015).

To consider the similarity between objects more precisely, or alternatively the distance between objects, a distance function/measurement needs to be defined. Several distance measures are discussed in the literature such as: Manhattan (L1-norm), Euclidean distance (L2 norm), Hamming distance (categorical attributes), and Minkowski distance (p-norm) (Provost & Fawcett (2013)).

In single and multiple imputations there are various methods available, one of which is based on matching and is commonly known as nearest neighbour imputation (NNI). It employs the K-Nearest Neighbours (K-NN) algorithm. When dealing with a database in which the data points are separated into several non-overlapping classes and the aim is to predict the classification of a

new point K-NN is employed as a non-parametric approach by which an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its K nearest neighbours (Schilling (1986a), Schilling (1986b), Bezdek et al. (1986), Chavez et al. (2015)).

In other words, at each value of X classify to the class that receives the largest number of classifications or votes. This family of classifiers are known as majority vote classifiers as follows: (Ju (2010), Tang & He (2015))

$$\mathcal{C}(y, \mathbb{S}) := \underset{c \in classes}{\arg\max}\ score(c, NN_k(y, \mathbb{S})) \tag{1}$$

Here $NN_k(y, \mathbb{S})$ returns the k nearest neighbours of instance y in space $\mathbb{S}$, *arg max* returns the argument (c in this case) that maximizes the quantity that follows it, and the majority scoring function is defined follows:

$$score(c, \mathbb{N}) = \sum_{y \in \mathbb{N}} \mathcal{I}[class(y) = c] \tag{2}$$

Here the expression $\mathcal{I}[class(y) = c]$ has the value one if $class(y) = c$ and zero otherwise.

When matching census and survey data, the main objective is to find the best match in the survey data for jth population household at the gth area with $\mathcal{T}(Y_{jg})$ attributes together with its individuals out of which the ith individual is defined with $\mathcal{T}(X_{ijg})$ attributes. Once this matched item is found, the aim is to impute extra attributes at the household level using the sample data. To achieve this goal we employ the K-NN approach discussed before.

The pseudo code for matching population- and survey-specific data at household- and individual-level for area g using the K-NN approach is presented in Algorithm 3. In this algorithm, for each population household (i.e. Y_{jg}), the set of most similar households in the sample data is found using majority vote classifiers. Then, we search among the individuals in the survey households within the feasible set of solutions (found in the previous step in the K-NN algorithm) and compare their characteristics with the individuals located in the target population household (i.e. Y_{jg}). The main goal of using this algorithm

is to find the most similar household in the survey considering the household- and individual-specific attributes.

Algorithm 3 K-NN Algorithm for Matching Population- and Survey-Specific Data at Household- and Individual-Level for Area $g \in \{1, \ldots, G\}$

Inputs:

(i) Individual- and household-level population data attributes: i.e. $\mathcal{T}(X_{ijg})$ & $\mathcal{T}(Y_{jg})$; $i \in \{1, \ldots, N_g\}$ & $j \in \{1, \ldots, M_g\}$

(ii)Individual- and household-level survey data attributes: i.e. $\mathcal{T}(x_{ijg})$ & $\mathcal{T}(y_{jg})$; $i \in \{1, \ldots, n_g\}$ & $j \in \{1, \ldots, m_g\}$

Output: Synthetic population of individuals and households with an extra attribute

Algorithm:

1: **for** $\forall\ Y_{jg} \in \mathbb{U}_g^{(H)}$ **do**

2: Calculate $\mathcal{C}(Y_{jg}, \mathbb{S}_g^{(H)}) := \underset{c \in \mathcal{T}(Y_{jg})}{\arg\max}\ score\Big[c, NN_k(Y_{jg}, \mathbb{S}_g^{(H)})\Big]$

3: Set the associated household unit (with their individuals) in the survey data in as feasible solutions; i.e. $\mathbb{S}_g^* = (\mathbb{S}_g^{(H)*}, \mathbb{S}_g^{(I)*})$

4: Calculate $\mathcal{C}(\mathbb{U}_{jg}^{(I)}, \mathbb{S}_g^{(I)*}) := \underset{y_{jg} \in \mathbb{S}_g^{(H)*}}{\arg\max} \Bigg[\sum_{x_{ijk} \in \mathbb{S}_{ijg}^{(I)*}} \mathcal{C}(x_{ijg}, \mathbb{U}_{jg}^{(I)}) \Bigg]$

5: **end for**

4. Wealth Index in Bangladesh

Although Bangladesh has shown a promising performance in achieving the Millennium Development Goals in recent years, the country is still struggling with the issue of poverty according to the 2011 Human Development Index (Klugman (2011)). The first of the Millennium Development Goals established by the United Nations is the eradication of extreme poverty and hunger. The estimated poverty rate at the divisional level in Bangladesh varies from 26.2 percent in Chittagong to 42.3 percent in Rangpur division on the basis of the recent 2010 Household Income and Expenditure Survey. At the district level, the rate varies from 3.6 percent in Khulna district to 63.7 percent in Kurigram district according to a recent poverty mapping exercise (Jolliffe et al. (2013)).

The proxy wealth index offers potential as a pragmatic and quick means of assessing poverty status. However, measuring relative wealth or living standards of people in developing countries such as Bangladesh presents many challenges

since income data are often not available. Some attempts to calculate wealth indices in Bangladesh are presented by Pitchforth et al. (2007), Gunnsteinsson et al. (2010), Tareque et al. (2010). The wealth index is based on household asset information via principal components analysis (PCA) and has been widely used in many country-level demographic and health surveys to measure inequalities in household characteristics (e.g. LeClere & Soobader (2000)). This index can be utilized as an indicator of household level wealth that is consistent with expenditure and income measures. The distribution of the estimated wealth index has zero mean and unit standard deviation. The estimation procedure permits greater adaptability of the wealth index in both urban and rural areas.

In the current study the 2011 BDHS data is used to measure the wealth index in different parts of Bangladesh. A nationally representative sample is drawn from the 2011 BDHS covering the entire population of Bangladesh based on a sampling frame of enumeration areas (EAs) listed in the 2011 Population and Housing Census given by the BBS. The EAs are considered as primary sampling units (PSUs) for the survey and consists of an average of about 120 households. A two-stage stratified sampling design has been used to select 600 EAs (207 in urban and 393 in rural areas) in the first stage with probability proportional to the EA size, and in the second stage a systematic sample of about 30 households was selected from each of the selected EAs. The ultimate sample consists of 17,141 households and 83,731 household members (NIPORT (2013)). For each district, Figure 4 presents the total population and the sample size of the 2011 BDHS. Sample sizes vary between districts from 30 households in Bandarban to 1425 households in Dhaka, with median of 228 households.

Figure 4: District-specific population and sample size

In calculating the BDHS wealth index, PCA is used for assigning weight values to the wealth indicator variables (e.g. household possessions, utility services, etc.). The full list of these variables used for Bangladesh is available in NIPORT (2013). The basic steps followed in calculating the wealth for a household are: standardized calculation of z-score for the indicator variables, calculating the factor coefficient scores (factor loadings), multiplying the indicator values by the loadings, and finally summing to produce the household's

index value. Only the first principal component is used to calculate the wealth index. The resulting sum is itself a standardized score with a mean of zero and a standard deviation of one (Rutstein & Johnson (2004)).

Quantiles of wealth indices are calculated based on the distribution of the whole population of individuals rather than on the distribution of households. The distribution is population-based because it is thought that most analyses are concerned with poor people rather than poor households. To obtain the cut-off points, a weighted frequency distribution of households was constructed where the weights are obtained by multiplying the number of de jure members of the household by the sampling weight of the household. The distribution shows the national household population where each member is given the wealth index score of his or her household. The population is then ordered by the score and is divided into five groups each consisting of 20-percent of population. This provides a population weighted quantile for the index.

Sampling weights were calculated based on sampling probabilities specified separately at each sampling stage and cluster. This means that the probability of selecting a PSU from a stratum at the first stage and probability of selecting a household from the selected PSU are combined to calculate the sampling weight. These weights are needed to ensure the representativeness of the survey results at national and domain levels and are used to calculate the district specific estimations.

Details about how the district-specific SP of individuals and households is constructed for Bangladesh in the current study are discussed in Section 2. Once the SP of households and individuals are constructed, statistical matching is then performed by imputing the nearest possible records from the 2011 BDHS to estimate the wealth index for each household within the SP. The diagram presented in Figure 5 shows the steps followed to construct the district-specific SP of Bangladesh with estimated wealth indices for households. In areas with relevantly small sample sizes (e.g. Bandarban, Khagrachari, Naogaon, Rangpur, etc.) and where a close match can not be found for a specific household in the SP within the same district, the nearest neighbour algorithm will search for the

best match in similar districts. Such similarities are generally specified based on the geographic distance.

Figure 5: Diagram: population synthesis and statistical matching

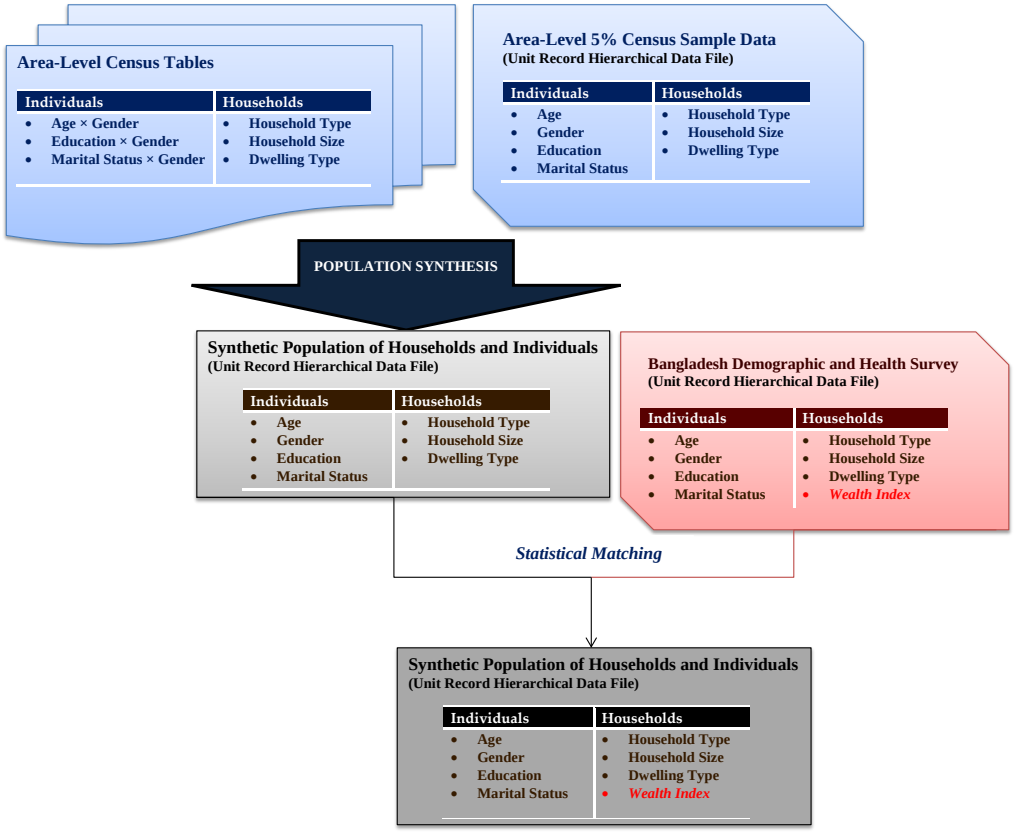
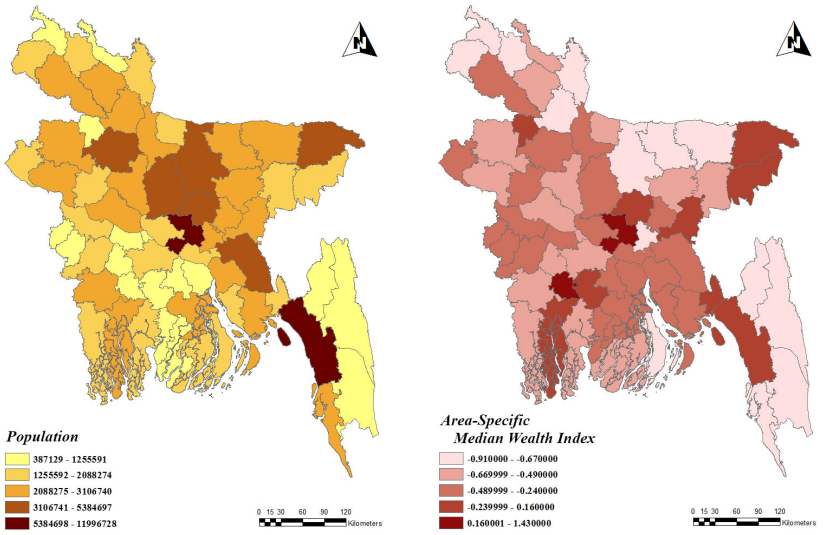


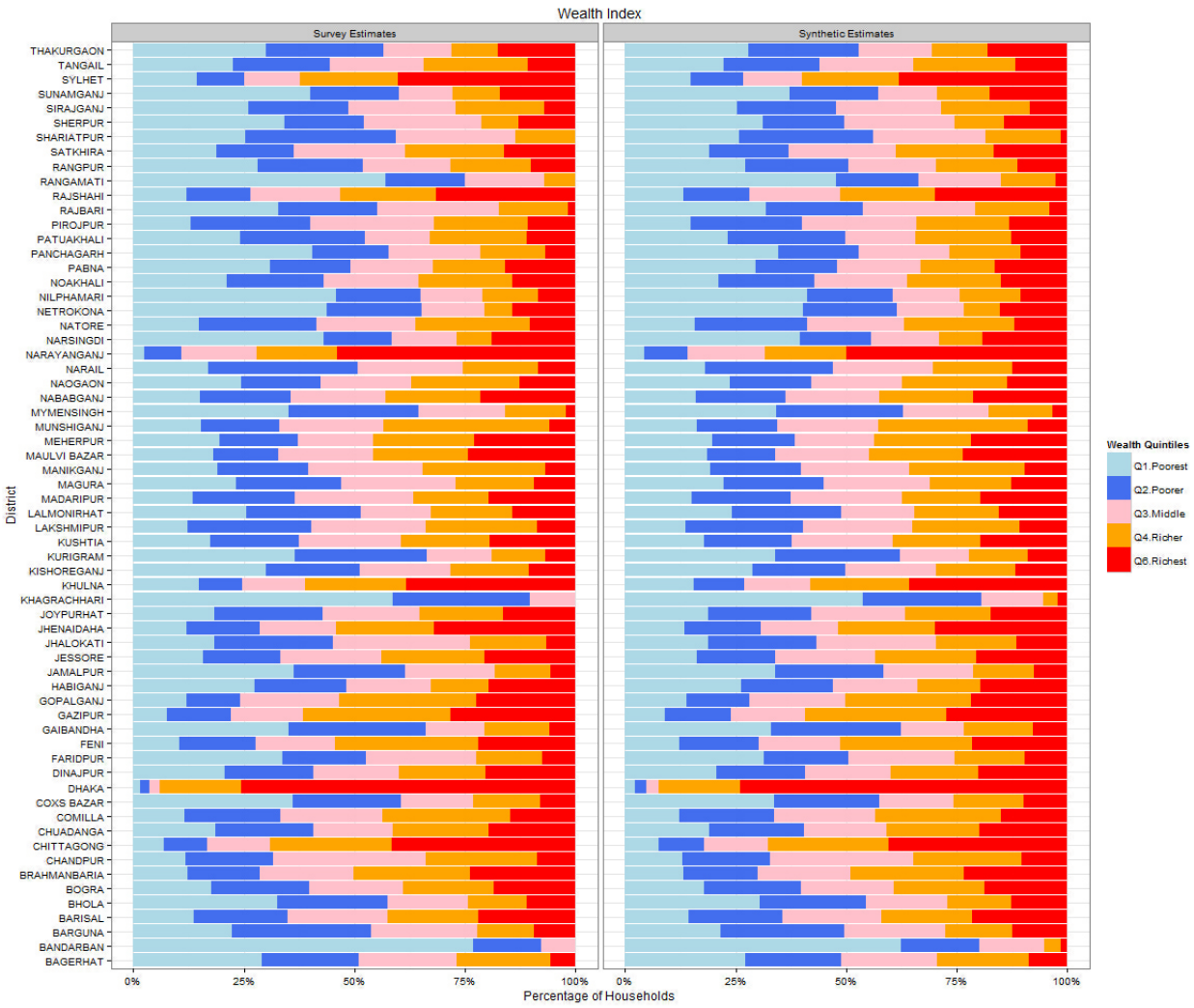
Figure 6 presents the district-specific population maps together with the median wealth indices estimated using the method presented in this paper.

Figure 6: District-specific population maps and survey estimates of median wealth index in Bangladesh



As mentioned before, the wealth index categorizes households into 5 wealth quantiles. The stacked graphs in Figure 7 show the percentage of population of each district in each of the five categories of wealth quantiles based on the direct Horvitz-Thompson estimators and the synthetic estimators obtained using the method presented in Figure 5.

Figure 7: Percentage of district-specific households in 5 wealth quantiles based on the direct Horvitz-Thompson estimators and the simulation-based synthetic estimators obtained based on the micro-simulation technique

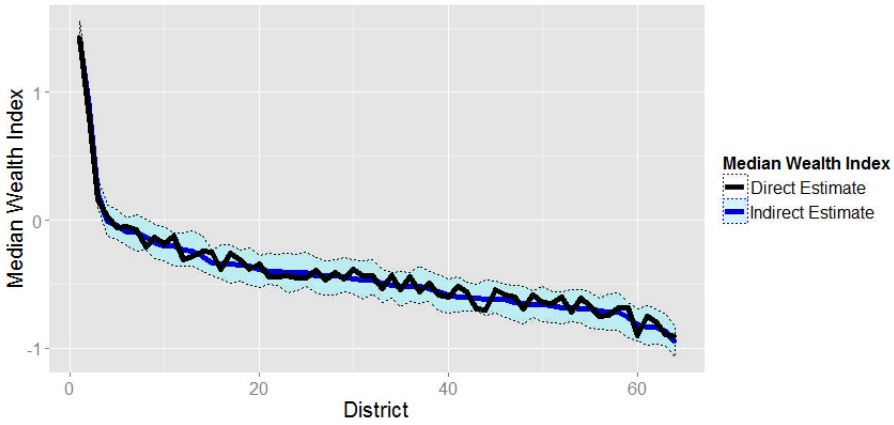


As previously discussed, the sample size in 2011 BDHS is relatively small for some districts. In areas with small sample sizes, sample units are mostly selected from the rural areas. As such, there are no households in some districts

categorized in the top categories of wealth indices. This is the case for districts such as Bandarban, Khagrachari and Rangamati. When dealing with such districts, the method presented in this paper could impute the wealth index from the closest match in similar districts and calculate the household- and district-level inferences more precisely.

When repeating the micro-simulation discussed in this paper based on which the area-specific wealth indices are estimated, it is possible to calculate a 95% simulation-based interval for the wealth index estimated for each area. Figure 8 presents the indirect estimation of median indexes for all the Bangladesh districts together with the 95% confidence intervals. The survey-based direct estimation of the median wealth index is also demonstrated in Figure 8. Note that the results are presented in Figure 8 in a way to be ascending for the area-specific median wealth indices estimated based on the micro-simulation technique presented in this paper. As can be seen in Figure 8, there are some differences between the direct and indirect estimates, as was expected from looking at Figure 7. That said, the direct estimates fall in the 95% confidence around the indirect estimates for almost all areas.

Figure 8: Direct and indirect estimation of area-specific median wealth indices for Bangladesh districts together with 95% confidence interval for the indirect estimates. Note that the results are presented in a way to be ascending for the indirect estimates for median wealth indices.



5. Discussion

Having fully disaggregated information about the population agents is needed for many micro level analysis tools such as microsimulation and agent based modelling. While in many countries census data is still the major source for geographically detailed estimates of populations and economies, such data is being released at higher levels of aggregation in the form of contingency tables, as releasing the fully informative disaggregated data while preserving confidentiality is very challenging. One way to overcome such a challenge is to use spatial-microsimulation techniques to generate a pseudo-census built by combining data sources, particularly surveys and censuses or surveys and administrative sources.

The literature on simulating artificial populations has undergone great development recently to help with spatially explicit and individual centred interaction models (micro-simulation or agent-based models). Such artificial populations are mostly derived using computer-aided simulation and are built from pseudo-census information obtained from anonymous survey and census information. A reliable SP aims at faithfully reproducing actual social entities, individuals and households, and their characteristics so as to represent reality as closely as possible. In the current study the Bangladesh census data is used to generate a district-specific SP of individuals and households.

In constructing the area-specific SP a step-wise approach is presented in this study which starts with a unit record file as a sample of the census using a smart scale-up algorithm. In the second stage of population synthesis and to correct the discrepancies between the cross-tabulations from the simulated SP in the fist stage and the census cross-tabulations, a heuristic algorithm is presented. A matching is then performed by imputing the nearest possible records among the Bangladesh's 2011 BDHS to estimate the wealth index for each household within the SP. While previous work has used spatial microsimulation techniques to derive an SP, this research extends these techniques by adding an imputation stage using a K Nearest Neighbour approach. This imputation stage is able to add data onto the SP from another survey which was not available in the original

SP. This is a significant advance in the method for deriving SPs. The results from our analysis show that the estimates calculated based on the technique presented in this paper are more representative than direct survey estimates which have limited sample for many small areas. This is mostly the case for the districts with some sample.

Census data is considered as the main source of statistical information for simulating area-specific synthetic populations of individuals and households together with their socio-demographic characteristics. This is while there are financial limitations for conducting census capable of producing reliable estimates of the corresponding population and improve erroneous census enumerations and census omissions at all required levels of geography. This is mostly the case in developing courtiers. That said, possible uncertainties in the census-based measurements will be projected in the simulated SP. This can be seen as the main limitation in the simulation presented in this paper as the Bangladesh census data is assumed to be the most accurate source of data based on which the SP is generated. It will be noted that, the levels of aggregation (of individuals) in the simulated SP follow the population structure given in the census data. Following Namazi-Rad et al. (2014b), a two-fold nested structure of the individuals and households within the target areas is considered in this paper to match the structure of Bangladesh census data. Other level of aggregation can be considered in population synthesis if sufficient information about associated population structures is provided.

Acknowledgment

The authors would like to thank Professor Ray Chambers for his insights to this study. The authors wish to gratefully acknowledge the help of Professor Pascal Perez in initiating a research collaboration through which this study was made possible.

References

Alderman, H., Babita, M., Demombynes, G., Makhatha, N., & Ozler, B. (2002). How low can you go? combining census and survey data for mapping poverty in south africa. *Journal of African Economies*, *11*, 169–200.

An, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, *40*, 151–189.

Anderson, R., & Hicks, C. (2011). Highlights of contemporary microsimulation. *Social Science Computer Review*, *29*, 3–8.

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating synthetic household populations: Problems and approach. *Transportation Research Record: Journal of the Transportation Research Board*, *2014*, 85–91.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.

Ballas, D., Clarke, G., Dorling, D., Thomas, H. E. B., & Rossiter, D. (2005). Simbritain: A spatial microsimulation approach to population dynamics. *Population, Space and Place*, *11*, 13–34.

Ballas, D., Clarke, G., & Wiemers, E. (2006). Spatial microsimulation model for rural policy analysis in ireland: the implications of the cap reforms for the national spatial strategy. *Journal of Rural Studies*, *22*, 367–378.

Barrett, S., Eubank, S., & Smith, J. (2005). If smallpox strikes portland. *Scientific American*, *292*, 54–62.

Beckman, R., Baggerly, K., & McKay, M. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*, 415–429.

Bezdek, J. C., Chuah, S. K., & Leep, D. (1986). Generalized k-nearest neighbor rules. *Fuzzy Sets and Systems*, *18*, 237–256.

Birkin, M., & Clarke, M. (2011). Spatial microsimulation models: A review and a glimpse into the future. In J. Stillwell, M. Clarke, & J. Stillwell (Eds.), *Population dynamics and projection methods. Understanding population trends and processes* (p. 193208). Springer volume 4.

Burden, S., & Steel, D. (2015). Constraint choice for spatial microsimulation. *Population, Space and Place*, .

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). M-quantile models for small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 47–69.

Chambers, R., & Pratesi, M. (2014). Small area methodology in poverty mapping: An introductory overview. In G. Betti, & A. Lemmi (Eds.), *Poverty and Social Exclusion: New Methods of Analysis* (pp. 213–223). Routledge.

Chambers, R., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, *93*, 255–268.

Chandra, H., Salvati, N., & Chambers, R. (2015). A spatially nonstationary fay-herriot model for small area estimation. *Journal of Survey Statistics and Methodology*, *3*, 109–135.

Chavez, E., M.Graff, G.Navarro, & E.S.Tellez (2015). Near neighbor searching with K nearest references. *Information Systems*, *51*, 43–61.

Coondoo, D., Majumder, A., & Chattopadhyay, S. (2011). District-level poverty estimation: A proposed method. *Journal of Applied Statistics*, *38*, 2327–2343.

Deming, W., & Stephan, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 367–484.

Deville, J., Sarndal, C., & Sautory, O. (1991). Raking procedures in survey sampling. *Journal of the American Statistical Association*, *86*, 87–95.

Edwards, K., Clarke, G., Thomas, J., & Forman, D. (2011). Internal and external validation of spatial microsimulation models: Small area estimates of adult obesity. *Applied Spatial Analysis and Policy*, *4*, 281–300.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*, 355–364.

Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*, 180–184.

Farooq, B., Bierlaire, M., Hurtubia, R., & Fltterd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263.

Farrell, N., Morrissey, K., & ODonoghue, C. (2013). Simulated model for the irish local economy. In K. Edwards, & R. Tanton (Eds.), *Microsimulation Methods and Models* (pp. 187–200). London: Springer.

Ferguson, N., Cummings, D., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D. (2006). Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, *437*, 209–214.

Fienberg, S. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, *41*, 907–917.

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PLoS ONE*, *5*, e8828.

Groves, R., & Couper, M. (1998). *Nonresponse in Household Interview Surveys*. New York; Chichester: John Wiley & Sons.

Gunnsteinsson, S., Labrique, A. B., West, K. P., Christian, P., Mehra, S., Shamim, A. A., Rashid, M., Katz, J., & Klemm, R. D. (2010). Constructing indices of rural living standards in northwestern bangladesh. *Journal of Health, Population and Nutrition*, *28*, 509–519.

Guo, J., & Bhat, C. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *2014*, 92–101.

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, *15*, 1–24.

Huynh, N., Namazi-Rad, M., Perez, P., Berryman, M., Chen, Q., & Barthelemy, J. (2013). Generating a synthetic population in support of agent-based modeling of transportation in sydney. In *Proceedings of 20th International Congress on Modelling and Simulation (MODSIM 2013)*.

Hynes, S., Farrelly, N., Murphy, E., & O'Donoghue, C. (2008). Modelling habitat conservation and participation in agri-environmental schemes: A spatial microsimulation approach. *Ecological Economics*, *66*, 258–269.

Hynes, S., Morrissey, K., O'Donoghue, C., & Clarke, G. (2009). A spatial microsimulation analysis of methane emissions from irish agriculture. *Journal of Ecological Complexity*, *6*, 135–146.

Jolliffe, D., Sharif, I., Gimenez, L., & Ahmed, F. (2013). Bangladesh-poverty assessment: assessing a decade of progress in reducing poverty, 2000-2010. bangladesh development series; paper no. 31. washington dc; world bank.

Ju, B.-G. (2010). Collective choice for simple preferences. In J.-F. Laslier, & M. R. Sanver (Eds.), *Handbook on Approval Voting*. Berlin Heidelberg: Springer.

Klugman, J. (2011). Human development report 2011. sustainability and equity: A better future for all. *Sustainability and Equity: A Better Future for All (November 2, 2011). UNDP-HDRO Human Development Reports*, .

Kokic, P., Chambers, R., & Beare, S. (2000). A spatial microsimulation model with student agents. *International Statistical Review*, *68*, 259–275.

Lago, L. P., & Clark, R. G. (2015). Imputation of household survey data using linear mixed models. *Australian & New Zealand Journal of Statistics*, *57*, 169–187.

LeClere, F. B., & Soobader, M.-J. (2000). The effect of income inequality on the health of selected us demographic groups. *American Journal of Public Health*, *90*, 1892–1897.

Lenormand, M., & Deffuant, G. (2013). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation*, *16*, 1–16.

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*, *40*, 227–253.

Lovelace, R., & Ballas, D. (2013). truncate, replicate, sample: A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11.

Lovelace, R., Ballas, D., & Watson, M. (2014). A spatial microsimulation approach for the analysis of commuter patterns: From individual to regional levels. *Journal of Transport Geography*, *34*, 282–296.

Lu, H., & Gelman, A. (2003). A method for estimating design-based sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics*, *19*, 133–151.

Molina, I., & Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*, 369–385.

Morrissey, K., Clarke, G., Ballas, D., Hynes, S., & ODonoghue, C. (2008). Analysing access to gp services in rural ireland using micro-level analysis. *Area*, *40*, 354364.

Morrissey, K., O'Donoghue, C., & Farrell, N. (2014). The local impact of the marine sector in ireland: A spatial microsimulation analysis. *Spatial Economic Analysis*, *9*, 31–50.

Mozumder, P., & Marathe, A. (2005). Implications of an integrated market for tradable renewable energy contracts. ecological economics. *Ecological Economics*, *49*, 259–272.

Namazi-Rad, M., Huynh, N., Barthelemy, J., & P, P. P. (2014a). Synthetic population initialization and evolution- agent-based modelling of population aging and household transitions. In H. Dam, J. Pitt, Y. Xu, G. Governatori, & T. Ito (Eds.), *PRIMA 2014: Principles and Practice of Multi-Agent Systems - Lecture Notes in Computer Science- Volume 8861*. Springer.

Namazi-Rad, M., Mokhtarian, P., & Perez, P. (2014b). Generating a dynamic synthetic population - using an age-structured two-sex model for household dynamics. *PLoS ONE*, *9*, e94761.

Namazi-Rad, M., & Steel, D. (2015). What level of statistical model should we use in small area estimation. *Australian & New Zealand Journal of Statistics*, *57*, 275–298.

NIPORT (2013). *Bangladesh Demographic and Health Survey 2011*. National Institute of Population Research and Training (NIPORT).

O'Donoghue, C. (2015). *Handbook of Microsimulation Modelling (Contributions to Economic Analysis)*. Dublin: Emerald Publishing.

Pitchforth, E., van Teijlingen, E., Graham, W., & Fitzmaurice, A. (2007). Development of a proxy wealth index for women utilizing emergency obstetric care in bangladesh. *Health Policy and Planning*, *22*, 311–319.

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*, 685–704.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Rao, J. (2003). *Small Area Estimation*. Wiley.

Rassler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer; Lecture Notes in Statistics.

Rosenbaum, P. (2002). *Observational Studies*. (second edition ed.). New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. (2006). *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rutstein, S. O., & Johnson, K. (2004). *The DHS wealth index. DHS comparative reports no. 6*. Technical Report Calverton: ORC Macro.

Schilling, M. F. (1986a). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, *81*, 799–806.

Schilling, M. F. (1986b). Mutual and shared neighbor probabilities: Finite-and infinite-dimensional results. *Advances in Applied Probability*, *18*, 388–405.

Smith, D., Clarke, G., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A*, *41*, 1251–1268.

Stephan, F. (1942). Iterative method of adjusting frequency tables when expected margins are known. *Annals of Mathematical Statistics*, *13*, 166–178.

Tang, B., & He, H. (2015). Enn: Extended nearest neighbor method for pattern recognition. *IEEE Computational intelligence magazine*, *10*, 52–60.

Tanton, R., & Clarke, G. (2014). Spatial models. In C. ODonoghue (Ed.), *Handbook of Microsimulation Modelling (Contributions to Economic Analysis, Volume 293)* (pp. 367 – 383). Berlin Heidelberg: Emerald Group Publishing Limited.

Tanton, R., Vidyattama, Y., McNamara, J., Vu, Q. N., & Harding, A. (2009). Old, single and poor: Using microsimulation and microdata to analyse poverty and the impact of policy change among older australians. *Economic Papers: A Journal of Applied Economics and Policy*, *28*, 102–120.

Tanton, R., Vidyattama, Y., Nepal, B., & McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society, Series A*, *174*, 931–951.

Tareque, M. I., Begum, S., & Saito, Y. (2010). Inequality in disability in bangladesh. *PLoS ONE*, *9*, e103681.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics*, *91*, 773–792.

Tomintz, M., Clarke, G., & Rigby, J. (2008). The geography of smoking in leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, *40*, 341–353.

Treiber, M., & Kesting, A. (2013). *Traffic Flow Dynamics*. Springer.

Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, *52*, 177–200.

Wilson, A., & Pownall, C. (1976). A new representation of the urban system for modelling and for the study of micro-level interdependence. *Area*, *8*, 246–254.

Wu, B., Birkin, M., & Rees, P. (2008). A spatial microsimulation model with student agents. *Computers, Environment and Urban Systems*, *93*, 440–453.

Ye, X., Konduri, K., Pendyala, R., Sana, B., & Waddel, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *Transportation Research Board - 88th Annual Meeting*. Washington, U.S.A.